GSLetterNeo Vol.107
2017年6月

大量データの関係性の変化をインタラクティブに見るツールの紹介

松原 伸人
matubara@sra.co.jp

◆ はじめに

大量のデータの中から、どのデータがどのデータと似ているかとか、関連するかを、テキストデータであれば TF-iDF などで類似度を計算したり、数値データであれば統計的に相関を数値で表したりできます。

関連度合いがどのように変化するかは、グラフにすることで変化を見比べることができますが、データ量が多くなり、関連の個数が多くなると、グラフだけで見ていくのは大変です。

◆ 大量データの探索

大量のデータの関連度合いの変化を、時間を辿って見ていけるような探索ツールを開発しています。

今回紹介するツールは、時系列データの、時刻ごとの関係と、データの集計値などで作られたグループを表示します。

図1は、プロトタイプシステムのスクリーンショットです。

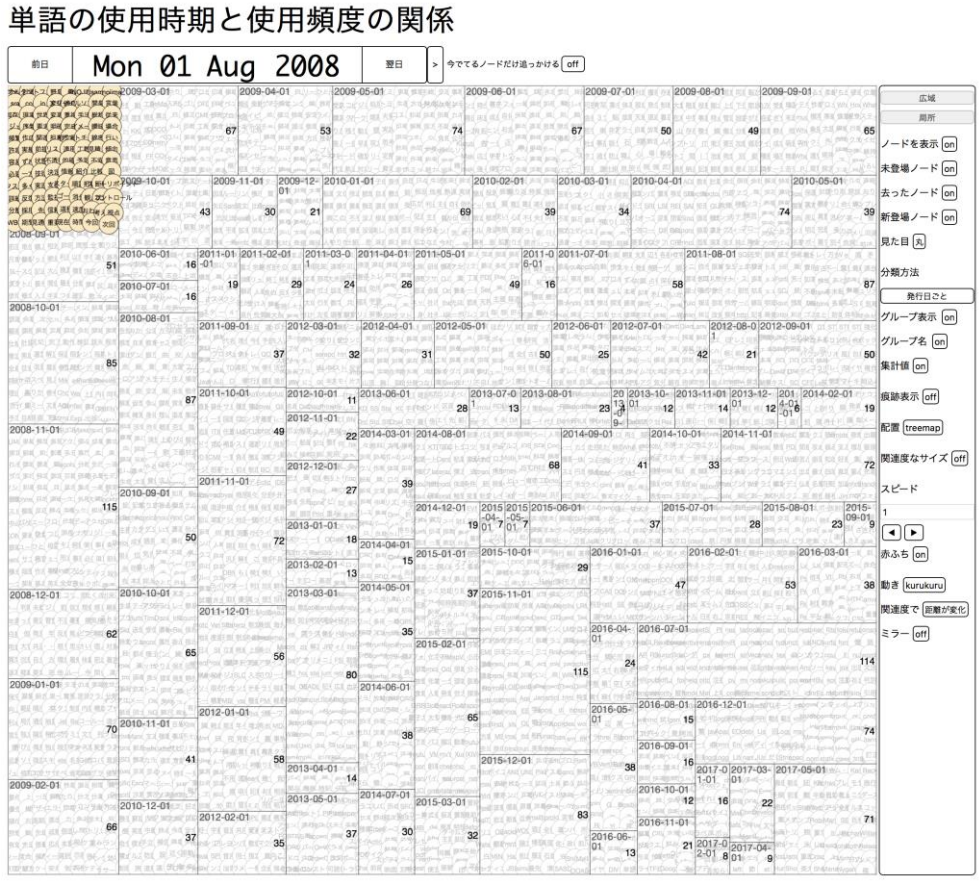

図 1 プロトタイプシステムのスクリーンショット

ツールを掲載するにあたり、GSLetterNeo の Vol.1 から Vol.106 のドキュメントに用いられているテキストデータ中の名詞の使用時期と使用頻度の関係を計算しました。

GSLetterNeo が発行された月ごとに形態素解析して名詞を抽出し、名詞が使われた時刻と、使用頻度を算出し、各月での使用頻度が近い名詞同士に関連を張って、関連度を記録してあります。

このプロトタイプシステムとデータは、KTL の GSLetterNeo ページに公開してあります。

[名詞の使用時期と使用頻度の関係]

(https://www.sra.co.jp/ktl/gsletterneo/index.html#words_tfidf_associations)

最初の状態は、GSLetterNeo Vol.1 が発行された 2008 年 8 月で登場した名詞を表示しています。

丸が名詞を表しています。黄色い丸は、選択している日に用いられていた名詞を表しています。

矩形のエリアは 2008 年 8 月から 2017 年 5 月まであり、それぞれの矩形の中にある名詞は、その日に初めて使用されたことを表しています。

日付を表す矩形エリアの右端の太字の数値は、名詞の数です。

「翌日」を押すと、次の号が発行された 2008 年 9 月に登場した名詞を表示します。

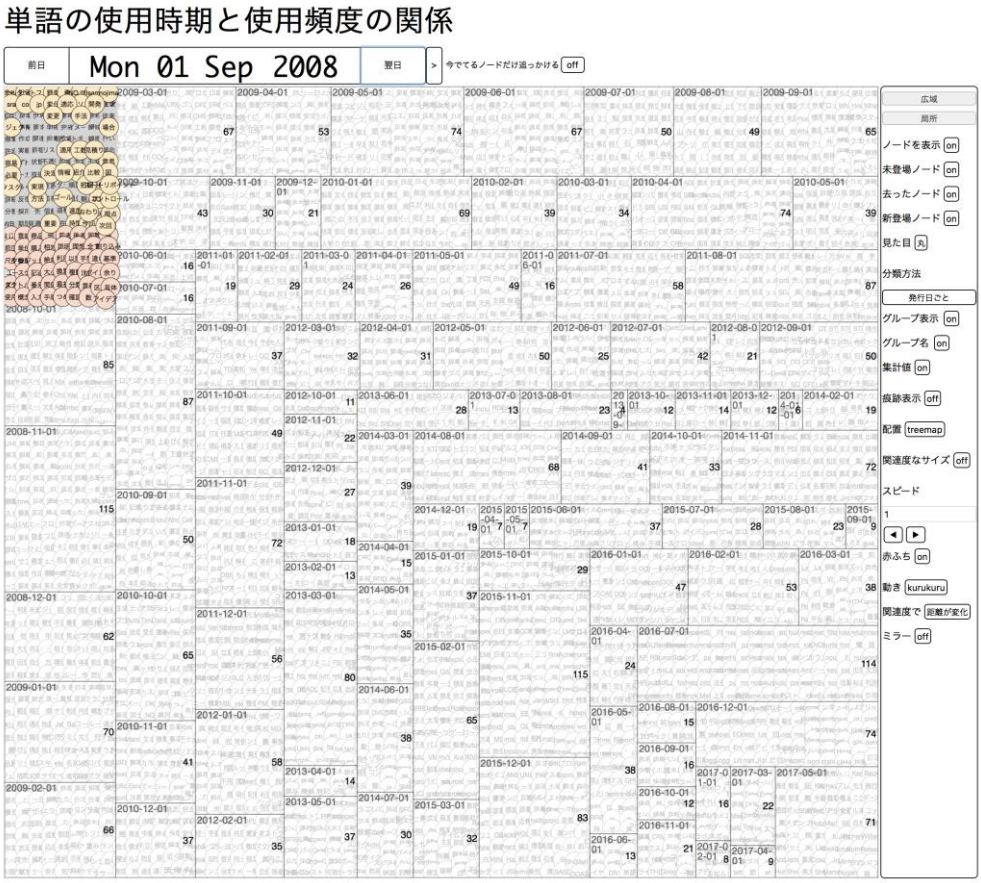

図 2　翌日ボタンを押した状態

薄い赤色の丸は、前日（直前に選んでいた日）には無かった、新たに登場した名詞を表しています。

灰色の丸は、直前に選んでいた日には使用されていたが、この日には使用されていない名詞を表しています。

灰色の丸は、日付を変えていくと少しずつ薄くなっていきます。10 回連続して使用されないと、色がなくなります。

日付を順に押していくことで、いつ頃にどのような名詞が使われていたのかを確認できます。

ノード(名詞)をクリックして選ぶと、選んだノードと関連するノードが、くるくると回ります。

選んだノードと関連度合いが高いほど、似たような動きをします。

例えば 2008 年 8 月の左上にある「アジャイル」を選ぶと、その日の名詞が全部回ります。創刊号で使われた名詞の使用頻度が同じためです。

「アジャイル」を選んだまま、「翌日」を押していくと、発行月ごとの「アジャイル」と使用頻度が類似する関連する名詞を順々に表示します。

「アジャイル」が使われていない日は、関連が無いので、くるくるする動きが止まります。

[ビデオ アジャイルと関連語]
(http://www.sra.co.jp/ktl/nobutomatsubara/gsletterneo_vol107_03.m4v)

右端のエリアには、ノードの表示や、上記で説明した新しいノードや去ったノードの表示の切り替え、データの分類方法、配置方法、関連を表す動き方の変更といったオプションを用意してあります。

GSLetterWeb 年表 (https://www.sra.co.jp/ktl/gsletterneo/index.html#chronicle) には、各号の特徴語を掲載していますが、これらは、TF-iDF が高い上位 10 名詞です。

これによると、2017 年 4 月において使用頻度が高い名詞である「年表」は、2016 年 2 月にはじめて使われて、同じように使用頻度高めの名詞には「矩形」「領域」「処理」「位置」「Orca」「Times」「Event」があることが見て取れます。

これらは全て年表化プログラム **orca** の仕組みを表す際に用いられている名詞で、「Orca」「Times」「Event」は年表の話題の中にしか出てない、GSLetterNeo シリーズ的に見ても、年表シリーズを特徴づける名詞だと思われますが、「矩形」「領域」「処理」「位置」は別々の日にバラバラに用いられていることから、わりと一般的に(GSLetterNeo の中では)用いられる名詞だと思われます。

試しに「Orca」「Times」「Event」でググってみると、全く関係のないページが出てきますが、「矩形」「領域」「処理」「位置」「Orca」「Times」「Event」の7つでググってみると、私の書いた記事そのものがトップに出て、続いて 2 つほど少し関係しそうな検索結果が並ぶと思います。

「矩形」「領域」「処理」「位置」は、わりと一般的に用いられていそうな名詞ですが、使用頻度が似ている「Orca」「Times」「Event」と一緒に検索に用いると、有効な検索が行えるようです。

◆ おわりに：応用と展望

今回紹介したツールは、自分が書いた記事や、ログ、自分が関わってきたことに関する記録などのような、私的なデータの時間的な変化と関係性を簡単に見られるようにすることを目指しています。

データの種類や解析方法が全く異なるデータでも、時間情報つきのグラフデータにできれば、様々なデータを同じように表示できます。

JSAI2017 にて **[時系列データ間の連関性と関係性理解のためのビジュアルインタラクティビティ]** (https://kaigi.org/jsai/webprogram/2017/pdf/822.pdf) というタイトルの論文を投稿しています。

この論文の中では、株価の変動の相関を見ることを事例として取り上げています。

興味のある方はこちらもご覧ください。

GSLetterNeo Vol.107
2017 年 6 月 20 日発行
発行者●株式会社 SRA 先端技術研究所
編集者●土屋正人

バックナンバを公開しています●http://www.sra.co.jp/gsletter
ご感想・お問い合わせはこちらへお願いします●gsneo@sra.co.jp

株式会社SRA
〒171-8513 東京都豊島区南池袋２－３２－８

夢を。Yawaraka Innovation
やわらかいのべーしょん